Table des matières

Table des figures:

REMERCIMENT

Nous tenons à remercier Monsieur Konstantin TODOROV notre responsable de formation et co-encadrent du projet pour sa patience, sa générosité, disponibilité et judicieux conseils mais également Monsieur Andon TCHECHMEDJIEV de par sa disponibilité, sa générosité dans le partage du savoir et sa réactivité.

INTRODUCTION

L'accès à l'information juste et équitable est devenu un problème majeur de nos jours avec l'essor du numérique et le boom des réseaux sociaux. Cette reconfiguration de l'espace public ouvre ainsi la porte à de nouveaux contenus, pas toujours vérifiés, ni toujours pertinents, pour le débat public, parmi lesquels certaines fake news. La propagation des fakes news a entrainé l'apparition du Fact-Checking. On retrouve un grand nombre de journaux qui disposent de structures particulières destinées aux tâches de Fact-Checking. Avec l'émergence de l'utilisation du Web au quotidien, des plateformes dédiées aux Fact-Checking ont été spécialement créées. Pour lutter efficacement et rapidement au fait de désinformés le publique, il a rapidement évolué auprès de nombreux grands médias. De nos jours, la pertinence attribuée à une information ne résulte plus d'une évaluation normative de son contenu par des experts mais émane plutôt d'une «agrégation numérique». C'est-à-dire que, désormais, les informations exposées sur la toile ne sont plus filtrées en étant au préalable passées au crible par des experts et journalistes. Elles sont à la place hiérarchisées a posteriori par des algorithmes de classement et de référencement qui dépendent en partie des clics et «likes» des internautes.

OBJECTIF DU PROJET

Le projet rentre dans le cadre du Travail d'Etude et Recherche (TER) et est partie intégrante du projet «ClaimsKG: A Knowledge Graph of Fact-Checked Claims» de Todorov, Tchechmedjiev et al.. Le projet global est réparti comme suit :

- un modèle claim data pour représenter les éléments de Fact-Cheking et les informations associées,
- un pipeline open source pour l'exploration et l'extraction de données à partir de sites Web de Fact-Cheking, et pour lever ces données en suivant le claim model,
- un KG (graph de connaissance ou Knowledge Graph) dynamique à grande échelle ouvertement disponible de revendications et de métadonnées associées, et
- une interface Web pour la recherche et l'exploration de la ressource.

Figure 1: Fact-cheking

Notre projet a donc pour but de faire de l'extraction des données à partir de sites Web de Fact-Cheking à l'instar de Pagella Politica qui est un site italien, Africa Check (le site est en français et traite les Fake News de l'Afrique) et Newtral qui est un site espagnol. Nous nous limitons donc aux Claims Extractor si bien qu'on a utilisé le Tagme pour le site italien.

Notez que le processus d'extraction est adapté individuellement à la structure de chacun des différents sites Web de vérification des faits, résultant en un ensemble d'extracteurs spécifiques. Les statistiques générées à chaque exécution du pipeline (globalement et par domaine) permettent de surveiller la fiabilité "des données extraites en détectant les problèmes potentiels qui peuvent être liés aux changements de structures des sites Web respectifs de Fact-Cheking qui peuvent être produits entre deux tronçons du pipeline.

Ainsi, nos sites sont structurés de façons différentes et ont été choisis pour faire un traitement d'information pour ensuite pouvoir extraire les données et les stockés.

Dans un premier temps on parle de données à extraire, des différentes informations importantes existant dans les articles du site Web et qui seront intégrer dans le KG du projet final. Les informations extraites sont :

- les titres des articles et des claims,
- la date de publication de l'article,
- l'auteur de l'article et de la claim,
- la véracité,
- les tags,
- les liens des articles,
- la source des sites et
- le corps de l'article.

En dernier lieu après extraction, on utilise un DataFrame pour affichées les données et les convertir depuis JSON en CSV, des fichiers CSV qui vont être utilisé ultérieurement pour l'autre partie.

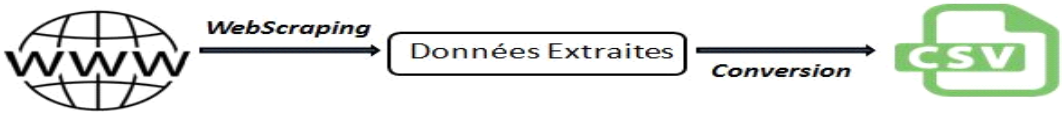

Figure 2: Scraping et Sérialisation

Partie I/ Fact-Checking

I.1/Definition

Il s'agit d'un terme anglais qui se traduit littéralement comme étant l'action de vérification. Le Fact-Checking désigne une forme de traitement journalistique qui vise à examiner et à vérifier les dires d'un responsable politique. Lors de débat public, le Fact-Checking est d'une grande importance afin de souligner la véracité, l'authenticité et la pertinence d'une information.

I.2/ Origine

Ce terme d'origine anglo-saxonne désigne une forme de démarche de vérification en interne des informations pour un organe de presse. Le Fact-Checking ne date pas d'hier. Dès son lancement, en 1923, le magazine *Time* avait déjà recruté une équipe de fact-checkers qui avait pour mission de vérifier scrupuleusement toutes les informations avant qu'elles ne soient publiées.L'exercice de la critique des médias se développe aux États-Unis à partir de 1988 avec la publication du livre d'Edward Herman et Noam Chomsky «La Fabrication du consentement». Les auteurs entendent démontrer comment les principaux médias interprètent les faits non pas de façon neutre et objective, comme leur obligé la déontologie, mais de manière tendancieuse, portés par le souci de véhiculer l'idéologie libérale (impulsée alors par le président américain Ronald Reagan et le Premier ministre britannique Margaret Thatcher) et du fait de la collusion entre les grands magnats de la presse et les hommes politiques. D'ailleurs, la mission Apollo a été récemment remise en question par de nombreux Américains persuadés qu'il s'agissait d'une intox.

I.3/Objectif du Fact-Checking

Deux champs distincts mais complémentaires sont visés:

- évaluer la pertinence et l'authenticité des propositions par les hommes politiques en confrontant avec des faits ou des discours anciens, dont on retrouve la trace via différentes sources d'information jugées fiables (agences officielles, instituts de statistiques, experts, etc.) «Pointeur du doigt les oublis, conférences à sens unique ou parfois même les inventions qui se glissent dans les discours».

- analyser le degré d'impartialité et d'objectivité des médias dans leur traitement de l'information, notamment sur les questions politiques, afin de vérifier si les faits ne sont pas instrumentalisés, de façon insidieusement tendancieuse afin de servir des intérêts partisans ou de dissimuler des conflits d'intérêts.

Dans les deux cas, sont signalés les imprécisions, inexactitudes et omissions (involontaires), les déformations de l'information (volontaires ou non), les informations vraies, mais présentées hors de leur contexte (volontaires ou non), les reprises d'informations fausses (volontaires ou non : on parle alors de désinformation), le mensonge (volontaire).

La vérification des faits demande des connaissances générales et la capacité des recherches rapides et précises (notamment, depuis l'affaire Fillon en 2017) dans le domaine juridique. Alors qu'à l'origine, ses praticiens étaient uniquement des journalistes intervenant dans le cadre de leurs enquêtes, aujourd'hui, n'importe qui peut retrouver des infos ou des vidéos en ligne, pointeur des mensonges, des contradictions, des raccourcis.

I.4 Les différentes étapes du Fact-Checking

- **Identifier le site Web**: Les rubriques «à propos» ou «mentions légales» sont particulièrement intéressantes à analyser car elles précisent généralement qui est l'auteur du site et quel est le cadre légal
- **Vérifier la fiabilité des sources**: Sont-elles indiquées (tout journaliste se doit de mentionner les sources des informations qu'il cite) ? Sont-elles légitimes, d'autorité, de notoriété publique, etc ?
- **Vérifier l'orthographe**: Un manque flagrant de correction orthographique ou syntaxique peut vous permettre de déceler l'usage d'un bot, par exemple, ou d'un traducteur automatique, etc.
- **Jetez un œil à la date de l'article**: Est-elle précisée ? Si oui, est-elle surannée ou récente ?
- **Si l'article est signé, vérifiez la fiabilité de l'auteur**: Généralement, les journalistes et blogueurs signent leurs articles.

- **Vérifier s'il s'agit d'un contenu dupliqué**: L'information apparaît-elle sur d'autres sites ? Il est intéressant de remonter jusqu'à **la source primaire**, celle qui a publié l'information en premier.
- **Valider la crédibilité de l'information**, ne serait-ce que simplement avec un peu de bon sens !

Partie II/Web scrapping

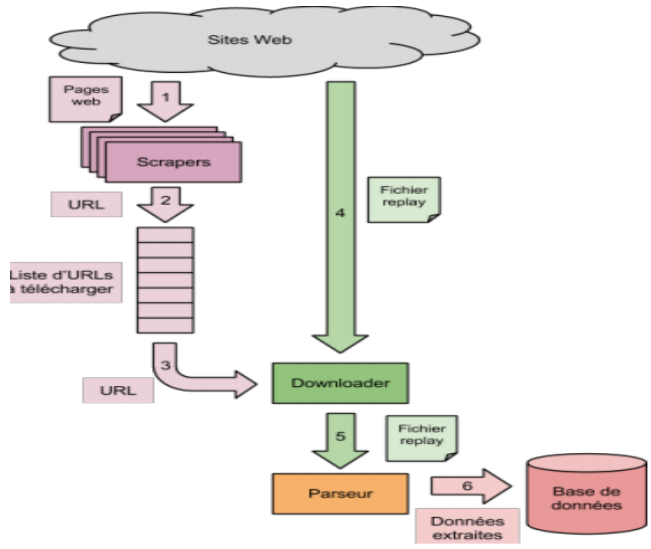

Figure 3: Processus de Scraping

II.1/Présentation de Web scraping

«Scraping» est un terme anglais signifiant littéralement «grattage». Autrement dit Scraper des données, c'est «gratter» des pages Web.Appliqué au Web, le terme, aussi connu sous le nom de Web scraping, Screen scraping, Web Data Mining, Web Harvesting, ou Web data extraction, renvoie à une technique d'extraction automatisée de contenu structuré pour stocker l'information voulue.

Recueillir des données sur le Web est parfois compliqué et quand cela est possible, il est difficile de pouvoir les télécharger. Cela revient à extraire du contenu d'une page Web, mais le scraping est en général assuré par des bots, ou robots, qui s'occupent de ce travail périodiquement. Ce dernier est une technique permettant l'extraction des données d'un site via un programme, un logiciel automatique ou un autre site. L'objectif est donc d'extraire le contenu d'une page d'un site de façon structurée. C'est une technique qu'il ne faut pas confondre avec le Web crawling, qui consiste, pour un logiciel, à scanner internet, à naviguer automatiquement de site en site pour collecter des données, dans un but d'indexation, facilitant ainsi la recherche de contenu, comme sur Google par exemple. Le but de cette technique est l'indexation, contrairement au scraping dont le but va être la récupération pure et simple afin de proposer le même contenu sur sa plateforme.

Le scraping permet ainsi de pouvoir réutiliser ces données, récolté du contenu sur un site Web, qui ne peut être copié collé sans dénaturer la structure même du document. Le scraping est également différent de l'usage d'une interface de programmation applicative (API), permettant au site source de contrôler le transfert des données aux tiers ré-utilisateurs en fournissant un accès gratuit ou payant.

Le Web scraping détient des éléments qui doivent être compris qui sont les suivants :

- Un site internet est constitué de pages Web, accessibles par des hyperliens dans d'autres pages Web.
- Une page Web est un fichier HTML retourné par le serveur, et associée à du JavaScript et des feuilles de style CSS. Ces fichiers HTML (.js et .css) sont appelés fichiers sources.
- L'information est entièrement contenue dans la page HTML que le serveur a envoyé. Pour accéder à cette information, il faut ouvrir le fichier HTML et la chercher dans les balises (utiliser XPath ou CSS).

II-2/Principe du Web scraping

Le scraping se fait en deux étapes : le téléchargement, du code HTML de la page à scraper, et son parsing. Pour obtenir le contenu de la page Web (téléchargement) il suffit de faire une requête et HTTP et d'attendre la réponse.

Obtention du code source (Téléchargement)

Premièrement : émettre une requête HTTP avec la fonction get de requests. À ce niveau du script on a le code source de la page sous forme de chaîne de caractère (str) dans la variable source. Ceci, si la requête a été effectuée avec succès. C'est-à-dire la réponse à la requête a pour code 200.

Récupération d'informations (Parsing)

Maintenant qu'on a tout le code source de la page, il ne nous reste plus qu'à récupérer les informations qui nous intéressent. Pour cela nous utilisons un Parser de code HTML. Il en existe plusieurs en Python.

Le Parser étant basé sur les balises HTML, on nous devons identifier les éléments qui correspondent aux informations qui nous intéressent. Pour faire cela, il nous faut inspecter la page dans notre navigateur. En faisant un clic droit sur la zone qui nous intéresse puis en cliquant sur "Inspecter l'élément". Dans la fenêtre qui s'affiche on peut voir la partir du code source de la page qui correspond à la zone qu'on a voulu inspecté. Le but est de trouver une caractéristique des éléments qu'on recherche, en se basant sur les attributs et les relations (imbrications) entre les éléments.

Figure 4: Web scraping des pages HTML et stockage des données sous différentes formes

II-3/.Extraction de données

Il est souvent utile de récupérer automatiquement des données à partir d'une page Web, en analysant le code HTML de la page pour extraire les informations qui nous intéressent.
Pour ce projet, on a extrait nos données en utilisant le langage interprété, multi-paradigme et multiplateformes Python, il favorise la programmation impérative structurée, fonctionnelle et orientée objet. L'extraction complète de données se fait par l'import différents modules (Beautiful Soup, requests, library, Tagme, etc).

3.1/ Beautiful Soup

Beautiful Soup est une bibliothèque Python qui utilise l'analyseur HTML / XML préinstallé et convertit la page Web / HTML / XML en une arborescence composée de balises, d'éléments, d'attributs et de valeurs. Cet arbre peut ensuite être "interrogé" en utilisant les méthodes / propriétés de l'objet Beautiful Soup créé à partir de la bibliothèque de l'analyseur.

Il permet :

- d'analyser une page Web pour déterminer le nombre de balises trouvées, le nombre d'éléments de chaque balise détectés et leurs valeurs. Vous voudrez peut-être les changer.
- de déterminer les noms et les valeurs des éléments afin de pouvoir les utiliser conjointement avec d'autres bibliothèques pour l'automatisation de pages Web, telles que Newtral, Pagella Politica, Africa Check.
- de comprendre la structure de la page Web, bien que nous utilisions d'autres bibliothèques pour effectuer l'acte de transfert, de transférer / extraire des données affichées dans une page Web vers d'autres formats, tels qu'un fichier CSV ou une base de données relationnelle telle que SQLite ou MySQL.
- de savoir combien d'éléments sont stylés avec un certain style CSS et lesquels.

Cet objet possède une méthode `select` qui prend en paramètre un sélecteur CSS, et retourne la liste des éléments HTML vérifiant ce sélecteur.

On peut également accéder au premier élément d'un type donné, en utilisant le nom de balise comme attribut de l'objet document.

Généralement, les méthodes `.find` et `.find_all` sont utilisées pour rechercher l'arborescence, en donnant les arguments d'entretien vrai qu'il existe d'autres méthodes comme (`prettify` pour améliorer la visualisation du code).

Les utilisations courantes de l'objet Beautiful Soup incluent:

1. Recherche par classe CSS
2. Recherche par adresse de lien hypertexte
3. Recherche par identifiant d'élément, tag
4. Recherche par nom d'attribut. Valeur d'attribut.

3.2/.Extraction de données depuis nos sites Web

L'extraction se fait en accédant au code source de la page d'accueil de chaque site :

Pagella Politica

Ce site est un site italien et pratique le fact-cheking. Il a une structure assez simple et nous avons essayé d'extraire les éléments qui s'y trouvent en fonction des besoins du projet.

VERO = Vrai

C'eri quasi = Quasi Vrai

Ni = Faux

Pinocchio andante = Presque Faux

PanzanaPazzesca = Archi Faux

Figure 5: Le site italien de Pagella Politica

Figure 6 Code couleur de Véracité pour le site italien

L'image ci-dessus représente la page d'accueil du site et les codes couleurs renvoient à la valeur de véracité.

Si nous accédons dans l'article, nous verrons facilement nos propriétés comme le montre le graphique ci-dessous.

Figure 7: Propriétés a extraire pour le site Pagella Politica

Ici nous avons :

- La claim qui est entre les guillemets
- Le titre de l'article en caractère plus grande accompagné de sa véracité avec la couleur verte qui signifie vrai
- Le nom de l'auteur de la claim
- La date de publication de l'article et la date de la déclaration de la claim
- La source de l'article également.

Tous ces éléments seront récupérés et stocké dans notre base de données.

Exemple d’extraction

L'extraction du nom de l'auteur de chaque article par exemple se fait comme suit:

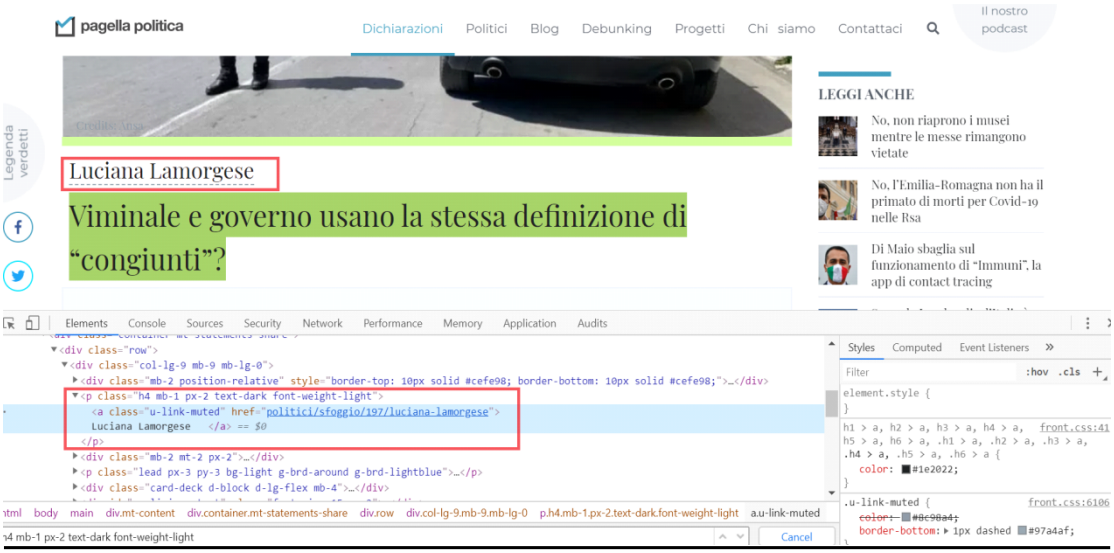

Figure 8: Inspection des balises

La code a été écris avec le langage de programmation Python en fonctionnel afin d’extraire le nom de l'auteur et le stocké dans un fichier JSON ou CSV .

Newtral

Newtral est un site espagnol de vérification des faits. Sa structure est différente de celle de Pagella Politica et à l'intérieur même des articles, il y'a des différences de structure.

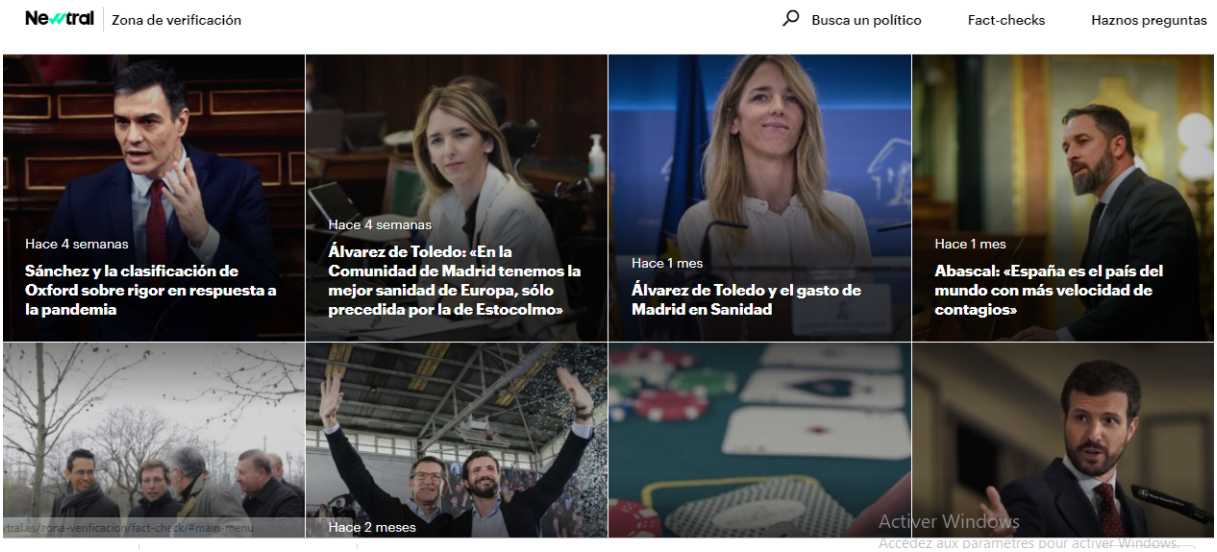

Figure 9: Page d'accueil du site espagnol Newtral

Comme vous pouvez le voir dans la page d'accueil il y'a parfois la claim, parfois le titre de l'article et pour pouvoir récupérer les éléments dont on a besoin, il faut rentrer dans l'article et mettre des règles spécifiques pour chaque article parce qu'il est possible d'avoir des articles avec plusieurs claims et plusieurs véracités. De même la place de la véracité peut varier d'un article à un autre, la également il faut trouver des règles spécifiques pour pouvoir les extraire.

Pour faire l'extraction dans ce site on a utilisé le langage de programmation Python en orienté objet .

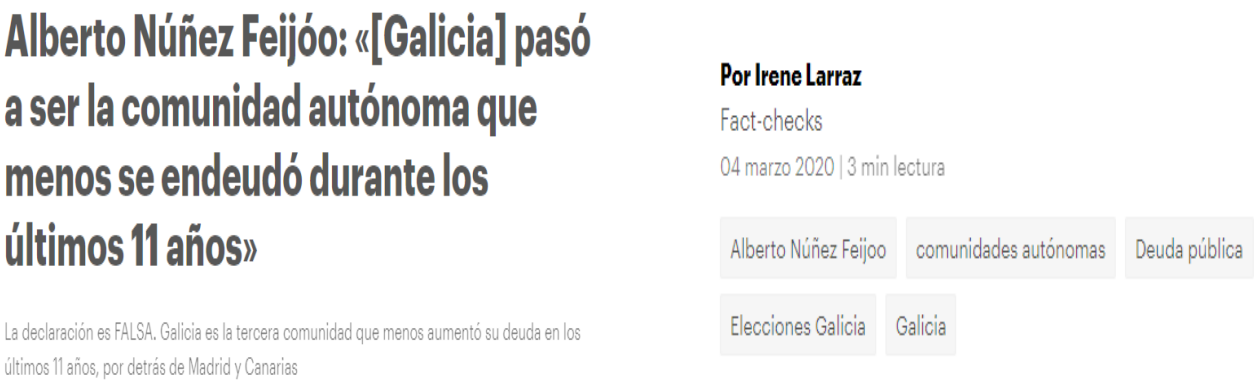

Figure 10: Extraction propriétés du site espagnol Newtral

Une fois dans l'article pour ce site, nous pouvons récupérer toutes nos propriétés. Seul la véracité a tendance à changer de position et peut être dans le texte (de même que les claims) ou dans l'image.

Pour les valeurs de vérité nous avons:

```
ENGAÑOSA, ENGAÑOSO : Trompeur
FALSO, FALSA, FALSOS : Faux
VERDADERO : Vrai
VERDAD A MEDIAS : A moitié vrai
```

Figure 11: Véracité pour le site espagnol Newtral

Africa Check

On a sélectionné ce site de Fact-Checking qui est francophone afin de traiter des langues différentes et la structure est également assez simple à l'image du site italien.

L'extraction de données dans ce site a été faite de la même manière que les autres sites, avec une structure de site différente et dont les propriétés sont directement visibles.

L'extraction de données sur ce site a été faite avec le langage de programmation Python en fonctionnel.

Figure 12: Site français Africa Check et ses propriétés

II-4.1/Les entités nommées

Le monde numérisé et connecté produit de grandes quantités de données. Analyser automatiquement le langage naturel est un enjeu majeur pour les applications de recherches sur le Web, de suivi d'actualités, de fouille, de veille, d'opinion, etc.

Les recherches menées en extraction d'information ont montré l'importance de certaines unités, telles que les noms de personnes, de lieux et d'organisations, les dates ou les montants. Le traitement de ces éléments, les «entités nommées», a donné lieu au développement d'algorithmes et de ressources utilisées par les systèmes informatiques.

La reconnaissance d'entités nommées est une sous-tâche de l'activité d'extraction d'information dans des corpus documentaires. Elle consiste à rechercher des objets textuels, c'est pour cela qu'on a utilisé l'outil Tagme dans notre projet pour présenté ses entités nommées et leurs fonctionnement.

4.1.1/Tagme

Tagme est actuellement l'un des meilleurs outils de liaison physique de la communauté scientifique, avec de très bonnes performances, en particulier lors de l'annotation de textes courts (c'est-à-dire ceux composés de dizaines de termes).

En termes simples, cet outil résout principalement le problème d'annotation de concept dans un morceau de texte. Dans tout morceau de texte, il est nécessaire d'extraire des concepts pertinents pour analyser l'ensemble du texte. Par conséquent, les concepts sélectionnés doivent satisfaire le concept. "Interrogeable" doit également répondre à "univoque", cet article doit s'assurer que le concept peut trouver la page correspondante dans Wikipedia, comme le concept "Réseau de neurones artificiels". En fait, c'est la définition plus formelle du concept "Réseau de neurones", c'est-à-dire le concept "Réseau de neurones" n'a pas sa page sur Wikipedia, contrairement au "Réseau de neurones artificiels", donc quand le concept "Réseau de neurones" apparaît dans l'article, nous voulons que la méthode le transforme en un "réseau de neurones artificiels". De plus, c'est sans ambiguïté. Par exemple, le concept "Apple" contient de nombreuses significations, y compris "pomme de fruit" ou "pomme". Pour savoir quelle est sa véritable signification à combiner avec son contexte, s'il fait

référence à Apple, alors le concept doit être traduit en “Apple Inc.”, le concept peut clairement exprimer le sens d'Apple.

Il existe différentes manières d’utiliser cet outil , dans l’un de nos site qui est le «Pagella Politica», nous avons utilisé l’API correspondante, il fallait faire l’import du package Tagme pour pouvoir faire l’extraction des entités nommés.

Pour faire des testes on a opté pour des outils de langage naturel qui sont le NLTK et spaCy .

4.1.2/NLTK

NLTK est une plate-forme leader pour la construction de programmes Python pour travailler avec des données de langage humain. Il fournit des interfaces faciles à utiliser à plus de 50 corpus et ressources lexicales telles que WordNet, ainsi qu'une suite de bibliothèques de traitement de texte pour la classification, la tokenisation, le stemming, le balisage, l'analyse et le raisonnement sémantique, des wrappers pour les bibliothèques NLP de qualité industrielle, et un forum de discussion actif .

Le traitement du langage naturel avec Python fournit une introduction pratique à la programmation du traitement du langage. Écrit par les créateurs de NLTK, il guide le lecteur à travers les principes fondamentaux de l'écriture de programmes Python, du travail avec les corpus, de la catégorisation du texte, de l'analyse de la structure linguistique, et d’autres fonctionnalités encore.

Figure 13:Traitement du langage naturel avec NLTK

4.1.3/spaCy

La reconnaissance des entités nommées (NER) apparaît comme une composante essentielle dans plusieurs domaines du NLP: résolution de coréférence, traduction automatique, recherche d'information, etc. Cette technique cherche à localiser et classer les entités nommées dans un texte en catégories prédéfinies.

Le traitement du langage naturel avec Python et spaCy vous montrera comment créer rapidement et facilement des applications NLP comme des chatbots, des scripts de condensation de texte et des outils de traitement des commandes. Cela permet de tirer parti de la bibliothèque spaCy pour extraire intelligemment la signification du texte, comment déterminer les relations entre les mots d'une phrase (analyse syntaxique des dépendances), identifier les noms, les verbes et d'autres parties du discours (balisage d'une partie du discours), et trier les noms appropriés en catégories comme les personnes, les organisations et les lieux (entité nommée reconnaissant).

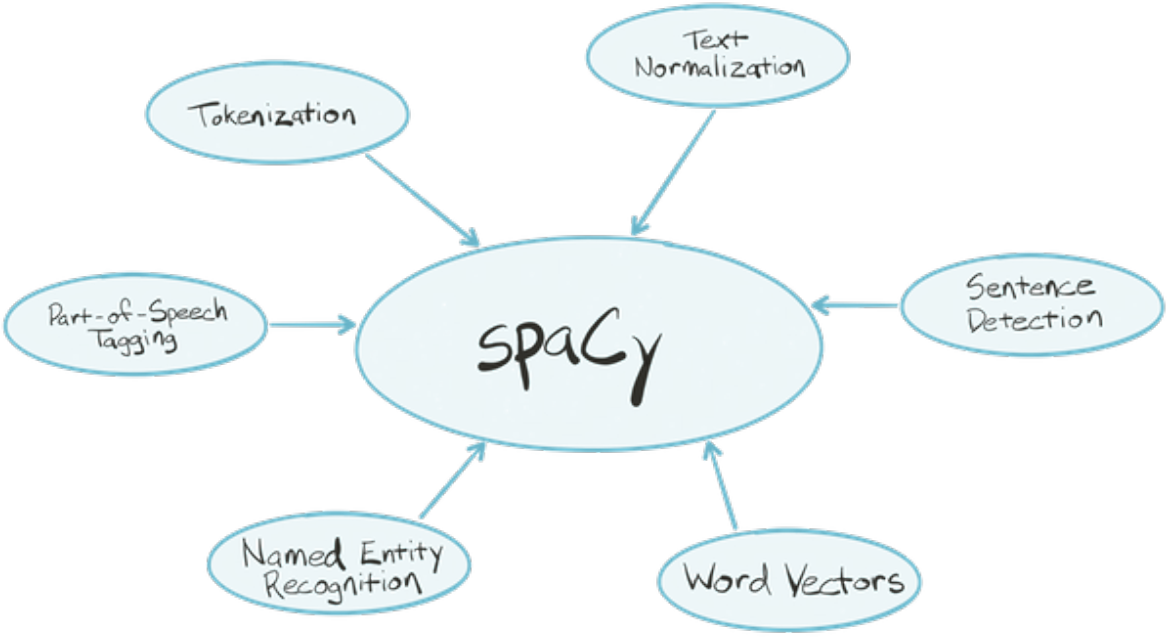

Figure 14: Traitement du langage naturel avec spaCy

II-5/ La réglementation pour le Web scraping

Les données en question sont majoritairement accessibles par un humain qui naviguerait sur la page Web. Face à la recrudescence des données, afin d'automatiser le processus, les scripts sont apparus.L'objectif d'un scraper n'est donc pas de voler des données mais de récupérer le plus souvent de façon périodique, les données d'un site. La masse d'information récupérée n'est pas illégale. Elle est même mise à disposition par le site.
Cependant, le scraping pourrait être considéré comme une pratique déloyale parasitaire. Ceci est particulièrement vrai lorsque le scraping est le fait d'un concurrent (cas où un site dans le domaine de l'énergie duplique les annonces postées sur un autre, par opposition à un comparateur de prix qui redirige les consommateurs vers la meilleure offre).

Ensuite, il est rare que le site copié ne parvienne pas à prouver avoir réalisé un investissement substantiel dans sa base de données. Un tel investissement permet habituellement de se prévaloir d'un droit de propriété intellectuelle. Cela dépend toutefois du type d'utilisation qu'en fait le scraper. Les conditions d'utilisation du site copié doivent être étudiées. Le droit pénal de la contrefaçon pourrait même s'en mêler.

Partie III/ Gestion du projet

III.1/Organisation interne

L'organisation interne s'est effectuée à travers des réunions hebdomadaires avec l'encadrant Mr Todorov et des échanges dynamiques avec Mr Andon.

Création d'un git pour mettre les ressources qu'on a besoin pour ce projet et les codes. La communication avec les encadrants se faisaient par mail en dehors des réunions et celle entre membre du groupe, nous organisons des réunions fréquentent sur skype afin de mieux avancer sur le projet malgré le confinement.

Utilisation de "Live Share" du Microsoft Visual Studio afin de mieux s'entraider.

Nous avons également récupérés les données du site italien ensemble et avons répartis des sites avec des approches différentes (fonctionnel et orienté objet)

III-2/Difficultés rencontrées

Certains sites changent régulièrement la structure du code HTML de leurs pages. S'il est possible que cela ne change rien visuellement dans un navigateur Web, cela va empêcher le programme d'en extraire les données. Bien entendu, le créateur du scraper pourra toujours le corriger pour prendre en compte les modifications mais il ne sera pas à l'abri d'autres changements ultérieurs. C'est le cas de Pagella Politica ou on avait réussi à récupérer tous les propriétés et faire le CSV, ils ont changé la structure de leurs sites 2 fois ce qui nous a retardé.

Il y'avait un souci pour le site de Newtral avec du JavaScript. On a pris un peu de temps pour comprendre comment récupérer les autres pages. Les sites étaient en format JSON et il y avait des informations qui manqués (La véracité, claims)

III-3/Diagramme de Grant

(A compléter :)

CONCLUSION

De nos jours, la pratique du Fact-Checking s'institutionnalise. Les discours politiques passent tous par une vérification méticuleuse des faits. Qu'il s'agisse d'information circulant dans les médias ou sur internet, tous se vérifient par des chiffres ou des citations. La pratique du Fact-Checking est vraiment essentielle pour que la population puisse profiter d'informations non erronées ou de discours non mensongers. En effet, Le Web a énormément influencé l'évolution du fact checking. Avec l'apparition et la domination de l'utilisation des réseaux sociaux, la forte circulation rapide des informations et la domination de la désinformation ont changé la pratique du Fact-Checking. Ce dernier est devenu incontournable face aux innombrables fausses informations et rumeurs à vérifier sur internet . Le projet a pour but de faire une etude sur le fact cheking , de choisir des sites et d'en extraire les données avec un langage de programmation qui est Python en fonctionnel et en orienté objet , utilisant des outils et des packages nécessaire pour faire l'extraction de données . ces dernières ont été enregistré dans des fichiers CSV pour qu'elles puissent être utilisé pour faire des graphes de connaissances .

BIBLIOGRAPHIE

Malo Gasquet, Darlène Brechtel et al. (2019). Exploring Fact-checked Claims and their Descriptive Statistics

Andon Tchechmedjiev, Pavlos Fafalios et al. (2019?). ClaimsKG: A Knowledge Graph of Fact-Checked Claims

Liens:

https://github.com/ImaneLamriou/TER
https://www.newtral.es/zona-verificacion/fact-check/
https://pagellapolitica.it/dichiarazioni/verificato
https://fr.africacheck.org/articles/
https://data.gesis.org/claimskg/
https://github.com/claimskg/claimskg-extractor